CLAIMS

What is claimed is:

1. A method of producing a linguistic dictionary, the method comprising:

storing explicitly substantially all orthographic variations of words in a finite state transducer database, and

storing, for each of the orthographic variations, a cut and paste code extended by a gloss code that indicates whether at least part of the orthographic variation should be converted between upper and lower case.

2. The method of claim 1, wherein the gloss code further indicates whether conversion should be performed between each single and double character sequence in the orthographic variation.

3. The method of claim 1, wherein the gloss code indicates one of (i)-(vii):

(i) Do nothing;

(ii) Convert first character to upper case;

(iii) Convert first character to lower case;

(iv) Convert word to lower case;

(v) Convert word to upper case;

(vi) Convert word to upper case and replace each single character sequence with equivalent double character sequence; and

(vii) Convert word to lower case and replace each double character sequence with single characters.

4. The method of claim 1, further comprising:

storing, for each word having an accented character:

a word having a composite form of the accented character; and

a word having an expanded form of the accented character that includes a base character and an accent character.

5. A linguistic dictionary comprising:

a finite state transducer database for storing explicitly substantially all orthographic variations of words,

wherein the database further stores, for each of the orthographic variations, a cut and paste code extended by a gloss code that indicates whether at least part of the orthographic variation should be converted between upper and lower case.

6. The linguistic dictionary of claim 5, wherein the extended gloss code further indicates whether conversion should be performed between each single and double character sequence in the orthographic variation.

7. The linguistic dictionary of claim 5, wherein the extended gloss code indicates one of (i)-(vii):

(i) Do nothing;

(ii) Convert first character to upper case;

(iii) Convert first character to lower case;

(iv) Convert word to lower case;

(v) Convert word to upper case;

(vi) Convert word to upper case and replace each single character sequence with equivalent double character sequence; and

(vii) Convert word to lower case and replace each double character sequence with single characters.

8. The linguistic dictionary of claim 5, wherein the database stores, for each word having an accented character:

a word having a composite form of the accented character; and

a word having an expanded form of the accented character that includes a base character and an accent character.

9. A computer program product comprising computer program means for performing substantially the steps of:

storing explicitly substantially all orthographic variations of words in a finite state transducer database, and

storing, for each of the orthographic variations, a cut and paste code extended by a gloss code that indicates whether at least part of the orthographic variation should be converted between upper and lower case.

10. The computer program product of claim 9, wherein the extended gloss code further indicates whether conversion should be performed between each single and double character sequence in the orthographic variation.

11. The computer program product of claim 9, wherein the extended gloss code indicates one of (i)-(vii):

(i) Do nothing;

(ii) Convert first character to upper case;

(iii) Convert first character to lower case;

(iv) Convert word to lower case;

(v) Convert word to upper case;

(vi) Convert word to upper case and replace each single character sequence with equivalent double character sequence; and

(vii) Convert word to lower case and replace each double character sequence with single characters.

12. The computer program product of claim 9, further comprising computer program means for storing, for each word having an accented character:

a word having a composite form of the accented character; and

a word having an expanded form of the accented character that includes a base character and an accent character.